Proceedings of Science and Mathematics
Faculty of science, universiti teknologi malaysia

https://science.utm.my/procscimath
Volume 24 (2024) 1-8

Understanding Viewer Opinions: Sentiment Analysis on Movie Review using VADER and LSTM Model

Ahmad Daniel Azahree, Norma Alias
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
Corresponding author: ahmaddaniel@graduate.utm.my

Abstract

Movies have become a significant tool for marketing and advertising that can influence consumer behavior. Reading film reviews is an important part of watching movie, as it helps viewers gain a general understanding of the film and provides filmmakers with feedback on their work is rated. Sentiment analysis, a method of determining whether a review has a positive or negative sentiment, plays an important role in this context. In this paper, we demonstrate and compare two different methods for sentiment analysis on textual IMDb movie reviews using VADER (Valence Aware Dictionary for Sentiment Reasoning) and LSTM (Long Short-Term Memory) model. VADER from NLTK module of Python and LSTM from extension of RNN will be used for this study. Both models will be assessed on the test set using evaluation metrics such as accuracy and precision for discovering which model predicts the sentiment analysis the best. The labelled data set comes from an online website called kaggle, which contains movie review information. Algorithms like the lexicon-based approach and the LSTM neural networks are trained using the chosen IMDb movie reviews data set. On the next work, LSTM will be used to classify the actual reviews from viewers of a few movies from IMDb website. LSTM are very effective for data like texts because it can relate to the context of the sentence very well. We preferred LSTM over VADER which help us predict our reviews better. We also can conclude that by using evaluation metrics or numerical indicators such as accuracy, precision, recall and F1 scores, we know that LSTM shows higher value than VADER.
.

Keywords: Machine learning; IMDb movie reviews; Natural language processing (NLP); Sentiment analysis; Long short term memory (LSTM); Deep learning

Introduction

Nowadays monitoring customer feedback and reviews is considered an important tool from a business perspective such as in the film industry. Filmmakers can create and present their masterpieces to the audience but getting timely reviews is a major input to planning the next business move. To overcome this issue the film makers need to analyse the movie reviews and identify the audience sentiment from the good and bad reviews. Sentiment analysis or opinion mining enable the filmmaker to analyse the reviews from social media websites automatically. Sentiment analysis also can be applied in many products or services. Therefore, developing automated tools for sentiment analysis is a big step in business.

Recently, deep learning has become a popular research topic in natural language processing. It represents an advancement over traditional machine learning by utilizing multiple layers of architecture to interpret data such as text, images, and sound. Deep learning is currently applied to various natural language processing areas, including text, voice, and speech, give impressive results. One of the key deep learning algorithms is the Long Short-Term Memory (LSTM) network. But, this approach is quite different from the lexical approach (VADER) where the words are mapped to the lexical dictionary corresponding intensities are obtained and then applied through a formula to obtain the overall sentiment score. Many researchers have applied automated sentiment analysis using the deep learning, the accuracy is much better than lexical-based approach. Based on current trend the automated sentiment analysis with good accuracy is really important to improve the movie quality that been produced.

We started with VADER, a lexicon-based approach. VADER uses a dictionary of sentiment words with their corresponding intensities, ranging from -4 to +4. VADER not only classifies reviews as positive or negative but also indicates the intensity of positivity or negativity in the text. However, VADER's limitation is that it can only detect sentiments of words included in its lexicon. New slang terms used in a review will not influence the sentiment classification, as these slang words are not part of the lexicon and therefore lack associated polarity scores.

LSTM, an extension of RNN, uses a machine learning approach to classify reviews as positive or negative. The key advantage of the LSTM model is its ability to understand context, which other neural network models may not achieve as effectively. Machine learning involves mathematical equations, so textual data must be converted into appropriate word vectors for use in these models. We used GloVe to obtain vector representations of words. These vectors are then fed into the models for training and prediction. This method differs significantly from the lexicon-based approach, where words are mapped to a lexicon, their corresponding intensities are retrieved, and a formula is applied to determine the overall sentiment score. The best performing approach could be for future prediction. Sentiment analysis is a widely used text classification tool that evaluates an incoming message to determine whether the sentiment is positive, negative or neutral. The primary application of sentiment analysis is classifying text into these categories. Study began with the work presented by [1], discussed various sentiment lexicons including VADER. A comprehensive overview of existing research was provided by [2] in their survey, which describes current techniques and approaches for opinion-oriented information retrieval. Early sentiment analysis research started with the foundational work that treated reviews as bags-of-words and aimed to classify them as positive, negative or neutral. A comprehensive survey of existing sentiment analysis methods [3], including an in-depth overview of sentence-level sentiment analysis found in the literature.

In this paper, we are focusing on the review analysis. This comes under the domain of text analysis and Sentiment analysis. Comparison of VADER and LSTM models (Lexicon based model vs. Machine learning model) for different evaluation metrics such as accuracy, precision, recall and f1 score. The dataset is collected by kaggle website; IMDb movie review.

VADER

The VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool doesn't have a straightforward mathematical equation like some machine learning models. Instead, it relies on a pre-built lexicon of words with associated sentiment scores and a set of grammatical rules to analyze the sentiment of a given text.

However, at a high level, the sentiment score (compound score) given by VADER can be thought of as a weighted sum of the individual sentiment scores of the words in the text. The sentiment scores range from -1 (most negative) to 1 (most positive), with 0 being neutral. The compound score might be calculated with:

$$\boldsymbol{CS = \frac{x}{\sqrt{x^2 + \alpha}}} \tag{1}$$

For equation (1), alpha is a constant typically set to 15, and x represents the sum of the polarity scores of all the words in a sentence. For example, ‘The movie is good and interesting’. In this sentence, ‘good’ and ‘interesting’ are sentiment words with polarity scores of 1.9 and 1.3, respectively. To calculate the compound score, find x, which is the sum of these polarity scores: $1.9 + 1.3 = 3.2$ and then, use the formula of compound score.

LSTM

LSTM serves as a machine learning model, necessitating the conversion of text into numerical vectors before feeding it into the model. To accomplish this, we employed NLTK tokenizer, assigning each unique word a sequence number. We padded this sequence to ensure uniform review lengths. Following preprocessing, we partitioned the processed data into distinct training and testing datasets. GloVe [4], coined from Global Vectors, is a model for distributed word representation. We used the GloVe file (GloVe.6B.50d.txt) which consisted of (400K words, 50d vectors). In this file, the first element of every line is a word and remaining elements are vectors for that word. We converted the data of this file to dictionary format where the key is a word and the value is a vector.

For model evaluation, we leveraged the Keras deep learning framework. The word vectors were passed to the LSTM embedding layer, serving as the initial layer of our network to obtain predictions.

Dataset

The dataset for Mobile reviews was collected from Kaggle [5]. The data utilized in this study for sentiment analysis is extracted from a data set imported from the online website https://www.kaggle.com/. The IMDb movie reviews data set contains 50K movie reviews. This data set contains 25,000 reviews, each categorized as negative or positive, and has three columns: id, sentiment, and review.

Having an equal number of positive and negative reviews in the data set offers several advantages for sentiment analysis performed by using machine learning techniques. It ensures balanced training, preventing biases towards either sentiment. The equal distribution facilitates accurate performance evaluation, allowing for reliable comparisons of different models. It improves model generalization by capturing underlying patterns for both sentiments. Additionally, it mitigates bias and promotes fair sentiment analysis results. Overall, the balance in the data set enhances the effectiveness and reliability of sentiment analysis models.

Pre-processing Data

i. Removing HTML tags: In order to obtain meaningful data, it is often very important to remove and clean the data of the data set by removing HTML tags, and other unwanted elements like the URL tags, hashtags and others which are of no use.

ii. Removing special characters: Removing special characters like punctuation marks, symbols, and numbers that have no use while performing sentiment analysis. This also includes converting all the characters into lowercase.

iii. Word Stemming: Stemming is the process of stripping a word of its suffixes and prefixes and returning it to its root form. In order to treat terms with the same root as the same word, the suggested system conducts stemming to lower the frequency of those words. Here, we have performed stemming using the NLTK (Natural Language Tool Kit) Python module.

iv. Removing stop words: This step involves removing stop words from the data. Usually stop words refer to words that add no meaning to text that is their presence does not have any use. Words like "him, "they", "it", "both", "how", "does" and others do have nothing to do with sentiment identification. So such stop words are removed which helps in decreasing the processing time of the model.

v. Tokenization: Tokenization is a data preprocessing technique of converting a separate piece of text into smaller parts like words, phrases, or any other meaningful elements called tokens which makes counting the number of words in the text easier. The proposed system performed tokenization at the word level so as to consider the sentiment polarity of each word.

vi. Word embedding: Machine learning algorithms usually do not have the capacity to interpret the data consisting of plain text or strings in its original form. They need numerical inputs in order to perform the tasks. So the process of mapping words from the lexicon to the corresponding vector of numbers, to derive words for sentiment predictions is called word embedding.

Evaluation/Performance Metric

Evaluation metrics are essential for assessing the performance of sentiment analysis models, ensuring that the models are accurate, reliable, and effective. Various metrics can be used to evaluate different aspects of the models. Below is an introduction to some commonly used evaluation metrics in sentiment analysis.

True Positive - Total number of reviews that the model has correctly predicted as positive**.**
True Negative -Total number of reviews that the model has correctly predicted as negative.
False Positive – Total number of reviews that the model has incorrectly predicted as positive.
False Negative- Total number of reviews that the model has incorrectly predicted as negatively.

Accuracy

The accuracy of a classification model can be defined as the ratio of total number of correct predictions made to the total number of predictions. The equation for accuracy can be given as (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Precision

Precision is defined as the ratio of the number of positive labels classified to the total number of positive labels. The equation for precision can be given as (3).

$$precision = \frac{TP}{TP + FP} \tag{3}$$

Recall

Recall is the ratio of true positives to true positives and false negatives. This is nothing but identifying the number of positives labelled correctly. The equation for the recall can be given as (4).

$$recall = \frac{TP}{TP + FN} \tag{4}$$

F1 Score

F1 score is used to summarize precision and recall in order to provide better results. F1 score can be defined as the harmonic mean of both precision and recall lying between 0 and 1. The equation for the F1 score is:

$$F1score = \frac{2 * precision * recall}{precision + recall} \tag{5}$$

Accuracy results

By determining the accuracy of each model, this performance metric is utilized to discover the most efficient algorithm from all of the selected algorithms. The accuracy of the two models are shown in table 1 below.

Table 1: Table of accuracy scores for both models

Model	Accuracy values
VADER	70.00%
LSTM	87.12%

Precision results

This performance metric depicts all possible positive predictions classified by the model. The selected models are evaluated considering this metric and the values of the precision scores of each model are shown in table 2 below.

Table 4.2: Table of precision scores for both models

Model	Precision scores
VADER	0.65
LSTM	0.89

Recall results

In order to find the optimized algorithm we use this performance metric. The below table 3, shows the recall values both model evaluated against the text data.

Table 3: Table of recall for both models

Model	Recall
VADER	0.86
LSTM	0.90

F1 Score results

F1 score is another performance metric that we have chosen in order to evaluate the selected models based on their performance. Table 4, shows the values of the F1 score for each algorithm.

Table 4: Table of F1 scores for both models

Model	F1 scores
VADER	0.74
LSTM	0.90

LSTM Model Performance

The LSTM model shows a minimal difference in accuracy scores between training and testing datasets, indicating its suitability for training approaches to data. Unlike some models that might show significant disparities between training and testing accuracies, the LSTM's ability to maintain a small margin between these scores suggests robust generalization to unseen data. This characteristic bring advantageous in tasks requiring the model to effectively learn from training data and generalize well to new, unseen instances, such as sentiment analysis.

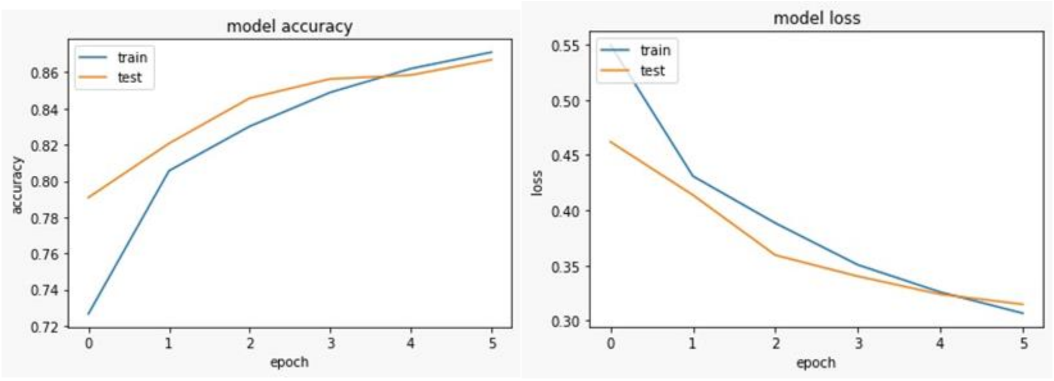


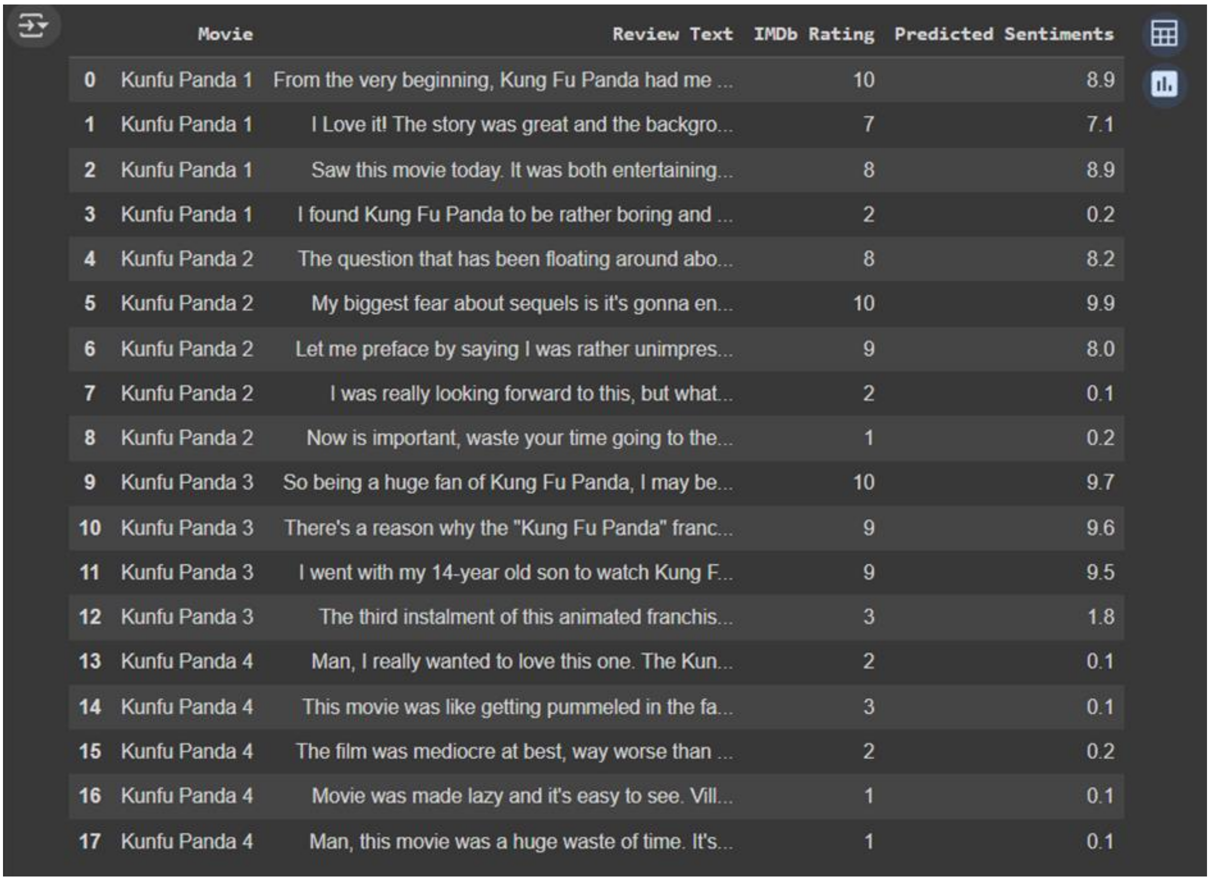
Figure 1 Graph plotted using the LSTM model to demonstrate the model’s accuracy

Sentiment Classifier Using LSTM Model

After conducting previous experiments, it has been established that LSTM models outperform VADER in sentiment analysis tasks. Consequently, for our analysis of IMDb movie reviews, specifically focusing on the 'Kung Fu Panda' film series encompassing all four installments.

We can observe that the predicted sentiment values closely match the IMDb results in figure 2. This demonstrates that the LSTM model provides accurate values, making it a reliable choice for sentiment analysis

.



	Movie	Review Text	IMDb Rating	Predicted Sentiments
0	Kunfu Panda 1	From the very beginning, Kung Fu Panda had me ...	10	8.9
1	Kunfu Panda 1	I Love it! The story was great and the backgro...	7	7.1
2	Kunfu Panda 1	Saw this movie today. It was both entertaining...	8	8.9
3	Kunfu Panda 1	I found Kung Fu Panda to be rather boring and ...	2	0.2
4	Kunfu Panda 2	The question that has been floating around abo...	8	8.2
5	Kunfu Panda 2	My biggest fear about sequels is it's gonna en...	10	9.9
6	Kunfu Panda 2	Let me preface by saying I was rather unimpres...	9	8.0
7	Kunfu Panda 2	I was really looking forward to this, but what...	2	0.1
8	Kunfu Panda 2	Now is important, waste your time going to the...	1	0.2
9	Kunfu Panda 3	So being a huge fan of Kung Fu Panda, I may be...	10	9.7
10	Kunfu Panda 3	There's a reason why the "Kung Fu Panda" franc...	9	9.6
11	Kunfu Panda 3	I went with my 14-year old son to watch Kung F...	9	9.5
12	Kunfu Panda 3	The third instalment of this animated franchis...	3	1.8
13	Kunfu Panda 4	Man, I really wanted to love this one. The Kun...	2	0.1
14	Kunfu Panda 4	This movie was like getting pummeled in the fa...	3	0.1
15	Kunfu Panda 4	The film was mediocre at best, way worse than ...	2	0.2
16	Kunfu Panda 4	Movie was made lazy and it's easy to see. Vill...	1	0.1
17	Kunfu Panda 4	Man, this movie was a huge waste of time. It's...	1	0.1

Figure 2 Average predicted sentiment by viewers

Conclusion

LSTM model outperformed the lexicon-based method in terms of accuracy and efficiency. To validate this conclusion, we evaluated the performance using metrics such as accuracy, precision, recall, and f1 score**.**

The LSTM model, exhibited superior performance in sentiment analysis compared to the lexicon-based method. LSTM's advantage lies in its ability to grasp contextual information and understand the nuances of language. By being trained on extensive text data, LSTM learns to predict words in a sentence based on their surrounding context, resulting in a highly accurate and robust model. On the other hand, the lexicon-based method relies on predefined sentiment dictionaries or lexicons. While it can yield reasonable results, it often struggles with contextual understanding and fails to capture subtle language nuances. Lexicon-based approaches typically assign sentiment scores to individual words or phrases and aggregate them to determine the sentiment of the entire text. This simplistic approach may overlook the complex interactions between words and their contextual meanings, leading to less accurate sentiment analysis.

To evaluate the performance of the two approaches, we utilized various metrics, including accuracy, precision, recall, and f1 score. Accuracy measures the overall correctness of sentiment classification, while precision and recall assess the model's ability to correctly identify positive and negative sentiments. The f1 score combines precision and recall, offering a comprehensive evaluation of the model's performance.

Based on these evaluation metrics, we consistently found that the LSTM neural network model outperformed the lexicon-based method in sentiment analysis accuracy. Its proficiency in capturing contextual information and understanding language intricacies enables it to achieve higher precision, recall, and f1 score values. Consequently, the LSTM model proves to be a more reliable and effective approach for sentiment analysis of IMDb movie reviews.

References

1. Hutto, C.J. Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
2. Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis." Foundations and TrendsÂő in Information Retrieval 2.1âĂŞ2 (2008): 1-135.
3. . Ribeiro, Filipe N., et al. "Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods." EPJ Data Science 5.1 (2016): 1-29.
4. Pennington, Jeffrey, Richard Socher, and Christopher Manning."GloVe: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
5. https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews.
6. Tarmizi, N., Saee, S., &amp; Abang Ibrahim, D. H. (2020). Detecting the usage of vulgar words in cyberbully activities from Twitter. International Journal on Advanced Science, Engineering and Information Technology, 10(3), 1117.
7. Sara Sabba, Nahla Chekired, Hana Katab, Nassira Chekkai, and Mohammed Chalbi. Sentiment Analysis for IMDb Reviews Using Deep Learning Classifier. In 2022 7th International Conference on Image and Signal Processing and their Applications (ISPA), pages 1–6, May 2022.
8. Cuk Tho, Yaya Heryadi, Iman Herwidiana Kartowisastro, and Widodo Budiharto. A Comparison of Lexicon-based and Transformer-based Sentiment Analysis on Code-mixed of Low-Resource Languages. In 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), volume 1, pages 81–85, October 2021